


Natural Language Processing

BSCS- 7th A

 Bahria University Discovering Knowledge	Department of Computer Science Bahria University, Lahore Campus
---	--

Twitter Sentiment Analysis

Name	Enrollment
MUHAMMAD ABDULLAH JAVED	03-134202-035
MUHAMMAD ALI RIASAT	03-134202-042

Department of Computer Sciences

Abstract

Twitter Sentiment Analysis project employs Natural Language Processing (NLP) techniques and machine learning models to classify airline-related tweets into positive, negative, or neutral sentiments. The project encompasses comprehensive data preprocessing, visualization, and machine learning strategies, including TF-IDF vectorization and handling class imbalance using SMOTE. Various classifiers are trained, and their performances are evaluated, emphasizing the significance of addressing imbalanced data. The outcomes contribute to a systematic approach for sentiment analysis, particularly in the context of the airline industry, with potential applications across diverse domains. We have achieved accuracy of 98%.

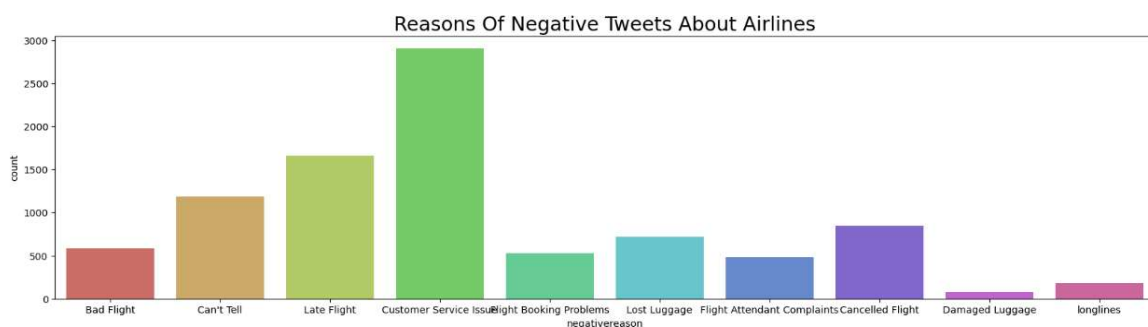
Data Cleaning and Preprocessing

Cleaning Steps

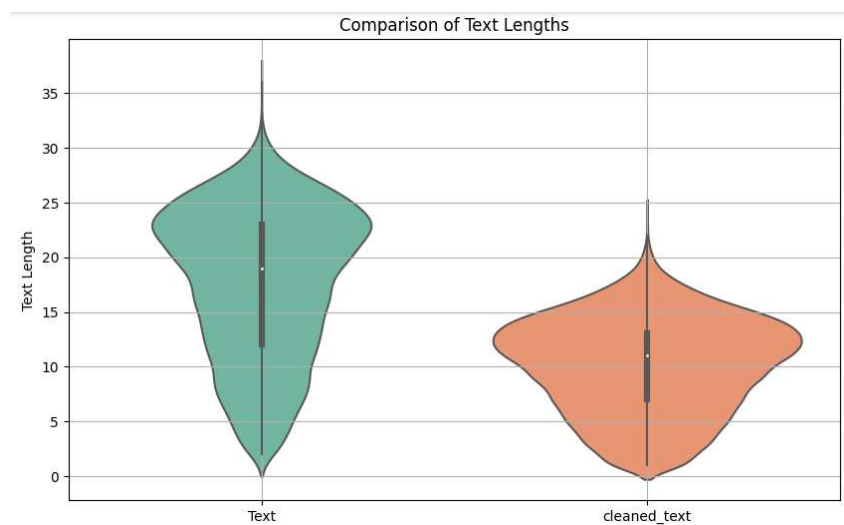
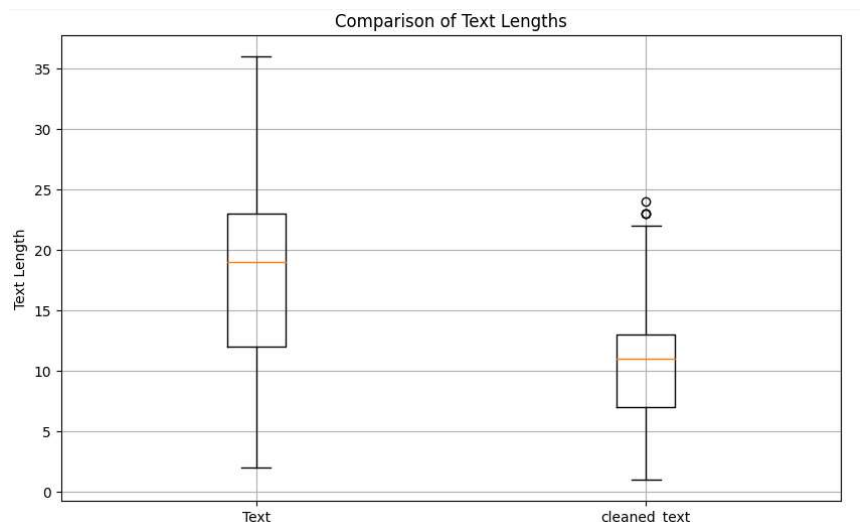
- Remove Username from Text.
- Remove HTML Tags.
- Remove Stop words.
- Remove URL from text.
- Remove Punctuation.
- Keep Character(A-Z) only in text.
- Separate Alphanumeric words and remove numbers.
- Re-complete words (like can't to cannot).

Stemming or Lemmatization somehow decrease the accuracy. So, we will ignore it.

One of the major issues in the dataset is with negative and neutral tweets. The models are predicting negative tweets as neutral. We have a column of “negativereason” in our dataset. We will concatenate the Negative reason with the text having negative sentiment. This will also help in improving accuracy.

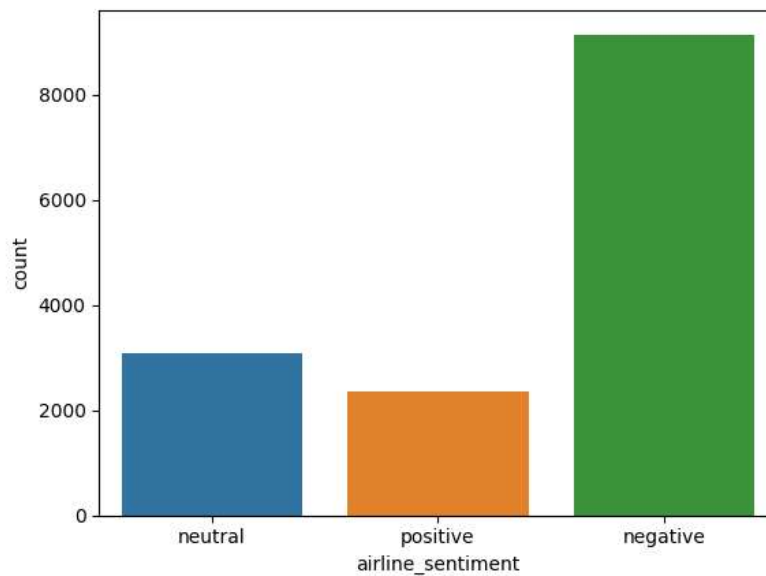


Text length Comparison (before processing and after)



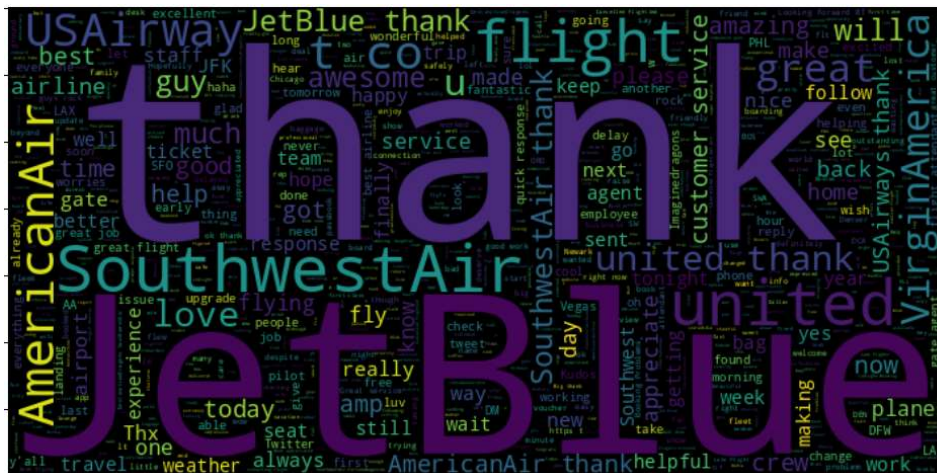
Tweets Count:

```
negative    9178
neutral     3099
positive    2363
Name: airline_sentiment, dtype: int64
```

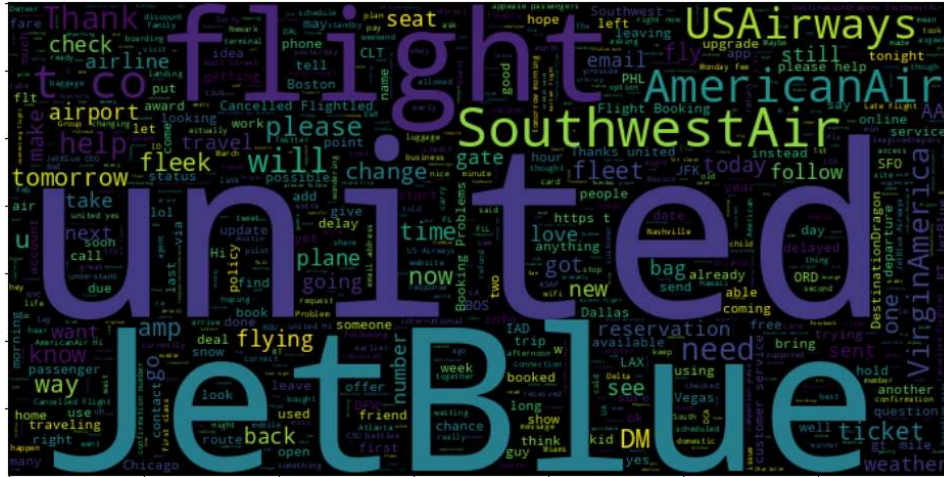


Word Cloud:

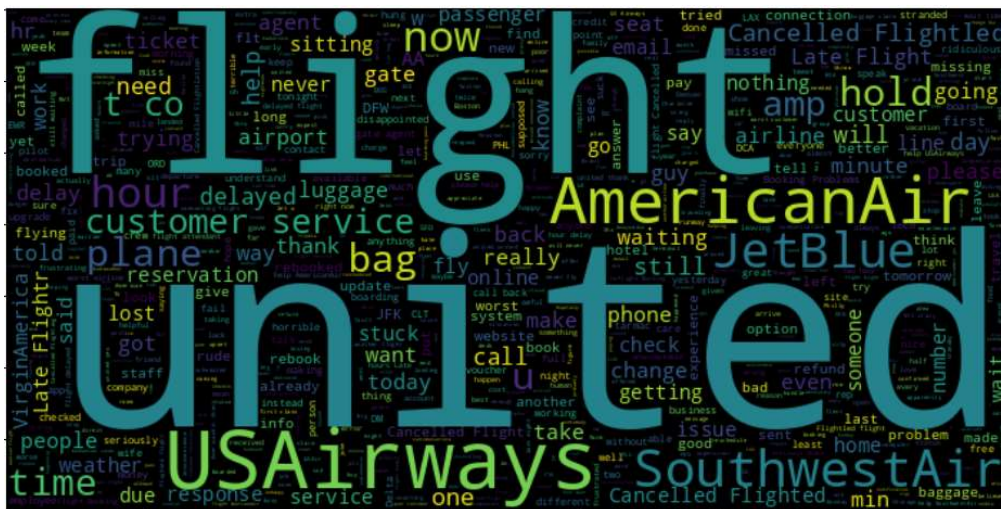
Positive:



Neutral:



Negative:



Handling Class Imbalance using SMOTE.

SMOTE stands for *Synthetic Minority Oversampling Technique*. SMOTE is an improved method of dealing with imbalanced data in classification problems. Imbalanced data is data in which observed frequencies are very different across the different possible values of a categorical variable. Basically, there are many observations of some type and very few of another type.

Machine Learning

TF-IDF Vectorizer

TF-IDF will transform the text into meaningful representation of integers or numbers which is used to fit machine learning algorithm for predictions. TF-IDF Vectorizer is a measure of originality of a word by comparing the number of times a word appears in document with the number of documents the word appears in.

Model Training

- Random Forest Classifier
- Support Vector Machine
- Naive Bayes
- Decision Tree Classifier

Model Evaluation

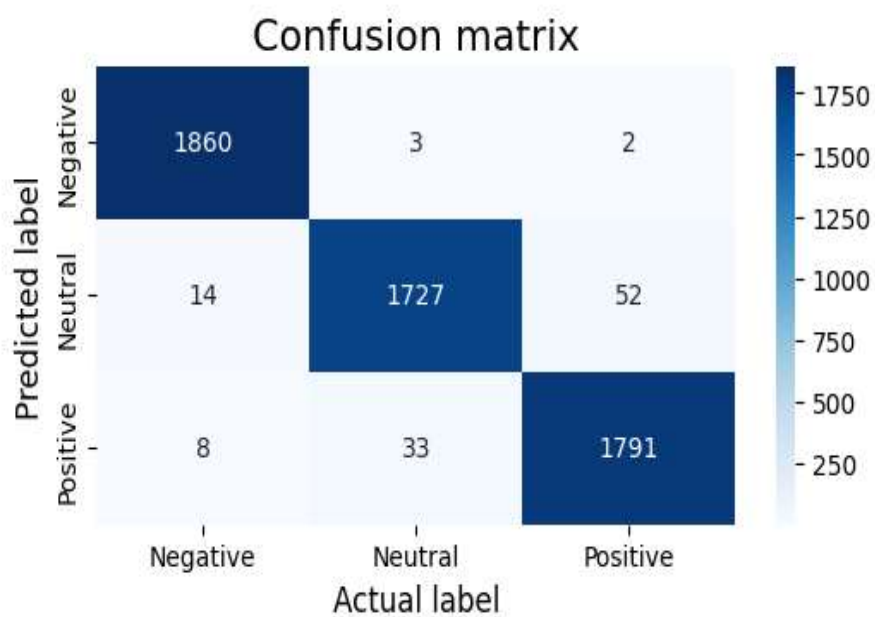
Train/Test Split	Model	Accuracy
Train Dataset = 70% Test Dataset = 30%	Random Forest Classifier	96.29%
	Support Vector Machine	97.58%
	Naive Bayesian	88.67%
	Decision Tree Classifier	93.92%
Train Dataset = 75% Test Dataset = 25%	Random Forest Classifier	96.48%
	Support Vector Machine	97.15%
	Naive Bayesian	88.34%
	Decision Tree Classifier	93.54%
Train Dataset = 80% Test Dataset = 20%	Random Forest Classifier	97.06%
	Support Vector Machine	97.95%
	Naive Bayesian	88.83%
	Decision Tree Classifier	94.26%

Classification Report:

Classification Report:

	precision	recall	f1-score	support
negative	0.99	1.00	0.99	1865
neutral	0.98	0.96	0.97	1793
positive	0.97	0.98	0.97	1832
accuracy			0.98	5490
macro avg	0.98	0.98	0.98	5490
weighted avg	0.98	0.98	0.98	5490

Confusion Matrix:



Conclusion

We have achieved best results using Support Vector Machine with Train-Test Split of (80,20). Accuracy of 98% approx.